DOCUMENT RESUME

ED 450 799 IR 058 067

AUTHOR Wharton, Sarah K.

TITLE The Role of Indexing in the Research and Development of Digital Libraries: A Call for Closer Examination of Domain-Specific Indexing and Thesaurus Construction To Improve Access to Digital Libraries.

PUB DATE 2000-08-00

NOTE 35p.; Master of Library and Information Science Research Paper, Kent State University.

PUB TYPE Dissertations/Theses (040)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Electronic Libraries; Indexes; *Indexing; Information Management; Library Automation; *Library Development; Online Searching; *Reference Materials; *Thesauri; Vocabulary

ABSTRACT

The purpose of this research is to investigate indexing issues that pertain to the development of digital libraries, including: the identification of sublanguage vocabulary; domain-specific indexing; and other indexing tools. Thesauri are needed for digital libraries in order to improve end-user access. To demonstrate the importance of thesauri to digital libraries, a mini-thesaurus will be constructed for the digital library domain. Specifically, the thesaurus will include terms that identify indexing methods and the infrastructure of digital libraries. The creation of the digital library thesaurus will reinforce the need for researchers to look at sublanguage vocabulary, and the need to use domain-specific indexing for all digital libraries. The terms will be extracted from abstracts and full text journal articles within three electronic research databases: "Library Literature," "Compendex," and from Internet resources. Three appendixes include a list of terms with definitions, the Digital Library Thesaurus hierarchical view, and the alphabetical view. (Contains 16 references.) (AEF)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 450 799

The Role of Indexing in the Research and Development of Digital Libraries: A Call For Closer Examination of Domain-specific Indexing and Thesaurus Construction to Improve Access to Digital Libraries

A Master's Research Paper submitted to the
Kent State University School of Library
and Information Science
in partial fulfillment of the requirements
for the degree Master of Library and Information Science

by

Sarah K. Wharton

August, 2000

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

D.P. Wallace

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

IR058067

BEST COPY AVAILABLE

Master's Research Paper by

Sarah K. Wharton

B.A., Cornell College, 1997

M.L.S., Kent State University, 2000

Approved by

Adviser Thomas J. Frosthech Date 8/15/00

Table of Contents

Introduction.....1-2

Literature Review.....2-3

Beginnings.....3

Architecture.....3-5

Subject Searching/Indexing.....5-6

Sublanguage Vocabulary.....6

Domain Indexing.....6-7

Vocabulary Issues.....7-8

Thesauri.....8-9

Other Indexing Tools.....9-10

Methodology.....10-11

Discussion.....11

Conclusion.....11-12

Works Cited.....13-14

Appendix A.....15-18

Appendix B.....19-21

Appendix C.....22-31

The Role of Indexing in the Research and Development of Digital Libraries: A Call For Closer Examination of Domain-specific Indexing and Thesaurus Construction to Improve Access to Digital Libraries

Introduction

Digital library (DL) research has become a complex and confusing undertaking because much of the research is conducted by people from a variety of academic disciplines including: computer science, library and information science, and the social sciences. The term "digital library" has a different connotation for researchers in the library field as opposed to those in the field of computer science (Levy 2000). Most librarians discern the word "library" to mean an institution which manages one or more collections; whereas computer scientists focus on the technology which builds and allows access to these collections. Even though there are competing views on what constitutes a digital library, its main goal must be to facilitate efficient and quick access to the information maintained within its site. Accessing information from a digital library may be difficult for the user because of the subject-specific vocabulary contained within the site. Hence, thesauri must be constructed for digital libraries in order to improve end-user access.

The purpose of this research was to investigate indexing issues pertaining to the development of digital libraries including: the identification of sublanguage vocabulary; domain-specific indexing; and other indexing tools. Thesauri are needed for digital libraries in order to improve end-user access. To demonstrate the importance of thesauri to digital libraries, a mini-thesaurus will be constructed for the digital library domain. Specifically, the thesaurus will include terms which identify indexing methods and the infrastructure of digital libraries. The creation of the digital library thesaurus will reinforce the need for researchers to look at sublanguage vocabulary, and the need to use domain-specific indexing for all digital libraries. The terms will be extracted from abstracts and full

text journal articles within three electronic research databases: Library Literature, Compendex, Inspec, and from Internet resources.

Literature Review

The researching and development of the digital library were conducted in a number of academic and research fields in the United States; but, because the information science field is understaffed with researchers, there are many aspects of its design that need to be more thoroughly examined. One such area of importance is indexing. Digital libraries were created with the idea of improving accessibility and usability; hence, in order to create effective DLs, one must examine the methods of indexing them.

There are many factors which make it virtually impossible to create an indexing standard for all digital libraries-- including the following: the human factor of subject searching vs. indexing; sublanguage vocabulary; and domain factors such as the role of domain in indexing (Bates 1998, 1185). It is sometimes assumed that subject searching and indexing is the same thing; however, the user's understanding of a particular subject area or topic may differ from the indexers'. For example, the user is trying to access information that he may or may not know. He must formulate a search query with words that he thinks will obtain relevant results. The indexer already has the record with which to work, and oftentimes a controlled vocabulary or sublanguage vocabulary. These terms or phrases may not be one with which the information seeker is familiar and hence if he or she does not use them, he or she will obtain poor results.

Sublanguage vocabulary refers to the vocabulary used in specific domains by people who work in that area. In order to improve access to users, researchers must examine the language used in individual digital libraries. By studying SLs, researchers hope to examine the structure of the domain the SLs represent in order to design databases and knowledge bases for that domain (Haas and He 1992, 721).

Domain factors are defined as those factors which characterize the domain. Examples of these factors are user's needs, terminology important to the domain, knowledge organization, and structure of the domain. All of these factors may influence how one indexes a particular digital library.

Beginnings

The development of the digital library occurred over a long period of time and is indebted to a variety of people (Kessler 1996,18). Computer scientists developed the science of storing and manipulating data electronically; and information scientists created techniques of representing, storing, and retrieving meaningful electronic data. The term "digital library" became popular in the early 1990s. When the World Wide Web was introduced, digital library access was redesigned to multiple remote services, which were WWW links (Kessler 1996, 23). Because of the explosion in the use of digital libraries, a Digital Libraries Initiative was established to maximize research potential.

The Digital Libraries Initiative was established in 1994 by the National Science Foundation (NSF), the Department of Defense Advanced Research Projects Agency (DARPA), and NASA. Some of the research focused on integration speech, natural language understanding, and interoperability technologies (Griffin 1998, 26).

Architecture

A solid digital library architecture has the following properties: it is decentralized, evolving, flexible, and it has a measure of scalibility--or control over the size of the library (The NSF/DARPA/NASA Sponsored University of Michigan Digital Library Project 1997). Since technology changes rapidly, the focus of the digital library cannot be centralized like a traditional library; but rather, it must be decentralized in the sense that it is not restricted by geography. The end-user must be able to access the information from a remote site.

The digital library must be made available to its user population; therefore, user access is a very important component in the design of the DL. Unlike a traditional library, there is no face-to-face meeting with a librarian to clarify problems with the system. The University of Michigan Digital Library project stresses the importance of an infrastructure such that intellectual work (finding, creating, and disseminating knowledge) is embedded in a persistent, structured context even though the underlying networked system is evolving.

Since the digital library does not collect books and other print materials, it does not rely on typical library cataloging tools such as MARC and AACR2. Digital librarians have identified three categories of metadata information that may be used to describe digital libraries: descriptive or intellectual, structural, and administrative (Tennant 1998, 30). Descriptive metadata identify the author of the resource, its title, and its subject heading. Structural metadata detail how an item is structured. The administrative metadata describe how a digital file was produced and its ownership. Because digital libraries needed an innovative approach to catalog their materials, the Dublin Core was used.

The Dublin Core is an ongoing effort by librarians, computer scientists, and others to create a standard that would describe a variety of objects within many different subject disciplines. It consists of 15 elements such as title, subject, and author. Describing data within digital libraries is difficult, and it is made even more difficult because of issues of interoperability.

Interoperability may be defined as the ability of multiple resources and multiple applications to interact (Kramer 1997, 125). One level of interoperability is the use of common tools and interfaces to provide uniformity for navigation and access. This process relies on human intelligence to provide any coherence of content; however, new technologies

such as personal digital assistants and nomadic computing models are being developed to help with this need (Lynch and Garcia-Molina 1996, 86). Lynch and Garcia-Molina believe that interface design is a key area for research. User behavior, needs, and objectives must be examined in order to improve digital library interfaces.

Another level is semantic interoperability, which is the ability of a user to access similar classes of digital objects and services distributed across heterogeneous repositories with mediating software. The Z39.50 standard was instituted to help support these different interactions.

Digital libraries must also be concerned with how user interfaces are linked to information manipulation/analysis tools. The problem confronting digital libraries and interfaces is how to incorporate intelligence into the system. The Digital Libraries Initiative has worked on the idea of integrating speech, images, and natural languaging into the systems. Now that the reader has a better understanding of the architecture of the DL, it is important to examine the issue of subject searching and indexing--since the main goal of the digital library is to make information available to a defined population of users.

Subject Searching/Indexing

On many occasions, searching and indexing are viewed as the same process (Bates 1998, 1186). However, searching may be defined as the "formulation of a query by the user"; and indexing may be described as "using words or phrases to describe a document." The main problem with searching and indexing is that the user often does not know what he is looking for; and the indexer's job is to index what is in the record (Bates 1998, 1187). The user may be at a disadvantage because he might not know the appropriate vocabulary, and the indexer is not consistent in his use of indexing terms throughout the indexing process. For example, indexer consistency studies have revealed that indexers have used a large variety of terms within subject description systems (Bates

1998, 1188). Hence, there is usually confusion on the user's part as to why he cannot find a document that he knows exists.

Sublanguage Vocabulary

Another indexing concern is sublanguages. Sublanguage (SL) vocabulary is the language used in a specialized domain or topic area by the people who work in that area (Haas and He 1992, 721). By studying SLs, it is the hope of researchers to find information about the language so that natural language processing systems can be designed more efficiently. In order to design the system more efficiently, a method to automatically identify the technical or domain terms in abstracts must be developed (Haas and He 1992, 731). The automatic identification of terms would aid in the retrieval of documents.

Also, it is essential to study SLs word-by-word and phrase-by-phrase for the following reasons: to identify the vocabulary of the SL; to identify the vocabulary that may be considered part of general language; and to determine the usage patterns that will create the grammar of the SL (Hass and He 1992, 721).

Domain Indexing

According to Marcia Bates, research in the design of content indexing (terms or vocabulary used for a specific subject topic) and access has revealed that domain-specific indexing provides users with better search results because it takes into account the users' needs, the terminology of the domain, and subdomain vocabulary (Bates 1998, 1200). Obviously, many subject domains use similar words in a different manner. In order to improve access to materials, the material must be indexed with its own specific vocabulary. Hjorland and Albrechtsen believe that the study of domain analysis in information science is extremely important. They argue that the best way to understand information science is to "study the knowledge-domains as thought or discourse communities, which are parts of society's division of labor"

(Bates 1998, 1200). If the designers of a digital library want to improve retrieval, then they must be concerned with domain indexing.

Vocabulary Issues

In a paper devoted to explaining the role of vocabulary in the field of information science, Michael Buckland proposed these ideas in connection with the use of the word vocabulary. He proposed that vocabulary can make digital libraries more cost-effective; and that vocabulary is the central component of digital libraries (Buckland 1999b). According to Buckland, vocabulary can make digital libraries more cost-effective, and increase returns on investment. Since enormous investments have been made in creating repositories over networks, Buckland rationalizes that any technique that will improve access to unfamiliar metadata will improve the rate of return on those investments. Buckland's second point is not as well defined. He contends that retrieval systems can have a series of sets or collections of information, and that "vocabulary" is an appropriate term for the variety of values in these sets (Buckland 1999b).

In order to support his ideas, Buckland and a group of researchers decided to examine in-depth the role of vocabulary in digital libraries. Because of an increase in subject-specific digital libraries, the researchers aim was to create an English language index to metadata vocabularies for those DLs. It was the hope of the researchers that the index would help users become more effective searchers by creating what they call an "Entry Vocabulary Module." This module would provide guidance in the transition from familiar vocabulary to unfamiliar vocabulary. In effect, it is a dictionary of associations between words found in titles and abstracts and the metadata vocabulary. According to Buckland, the Entry Vocabulary Module would support a ranked list of metadata terms for a portion of the text. This would then allow for computer-assisted categorization if text were submitted in a search query (Buckland 1999a).

Entry Vocabulary Modules are needed because of the following three reasons: more repositories are available over networks; there are more investments in providing indexing, categorizing, and other metadata; and with an increase in repositories, there are more unfamiliar metadata vocabularies (Buckland 1999a). Buckland's prototype Entry Vocabulary Module is web accessible at http://www.sims.berkeley.edu/research/metadata/oasis.html. The prototype includes English language indexes to BIOSIS Concept Codes, the INSPEC Thesaurus, and other sites.

There were other factors that Buckland's research considered including: the use of natural language processing techniques (Buckland 1999a). The natural language processing techniques were based on work by Ray Larson. Larson's "Classification clustering" utilized probabilistic interpretation of vector-spaced retrieval (Buckland 1999a). Originally, the Entry Vocabulary Modules were created from the frequency of occurrences of single terms. Currently, natural language parsing software identifies noun phrases and uses these phrases instead of individual words.

Buckland's work offers the following advantages: it provides an alternative to human crafting of links between vocabularies; it permits the searching of fragments within the metadata and databases; and it allows the user to choose search terms from a familiar vocabulary (Buckland 1999a).

Thesauri

There is a wide range of information systems (those systems that allow for information retrieval) with interrelated information, but how does one access this information in a timely manner? An article by Ralf Kramer discusses the possibility of integrating multiple thesaurus databases (Kramer 1997, 122). Digital libraries have a variety of information stored in them, and thesauri can make the retrieval of this information easier. A thesaurus maintains a uniform vocabulary for indexing documents within information systems. However, different information systems have subject-

specific vocabulary. Kramer suggests that in order to exchange information among other fields, a common thesaurus could be used in conjunction with the subject-specific thesaurus. A thesaurus federation would allow individual thesauri to maintain their autonomy, and they would be paid per thesaurus use. In turn, users would have access to general and specialized thesauri within this federation.

Other Indexing Tools

Other approaches to indexing digital libraries include the following methods: Centralized Indexing/Centralized Search; Distributed Indexing/Centralized Search; and Distributed Indexing/Distributed Search. Centralized Indexing is a type of method used by most current web indexing systems. The global information space is traversed on a periodic basis; and content is indexed by a central service. The Distributed Indexing method allows for customization to its local resource type, which is then transmitted to the indexing engine. In the Distributed Indexing/Distributed Search method, the indexing engines are distributed and have access to small sets of information repositories. The user search is distributed across a subset of indexing engines.

There are many problems associated with these indexing methods including: scalability, reliability, heterogeneity, and intellectual property. Scalibility refers to "scaling up" systems from small to large (Bates 1998, 1196). One question frequently asked by researchers is, "Will the system function as well when it is made larger?"

Also, how does one contend with the human vocabulary issue? That is, when a digital library grows, the number of words a human knows usually will not grow as well (Bates 1998, 1197). Hence, the number of documents retrieved will increase. One must also be concerned with keeping the indexed information reliable and continually updating it, which could be both time-consuming and expensive.

With respect to heterogeneity issues, digital libraries demand a wide range of information to be indexed; hence, user interfaces that will permit the user to see these varied resources must be used. One other concern with indexing the digital library is intellectual property issues. Some resource holders may be unwilling to release indexing information in any format.

There are two other approaches to indexing digital libraries. One approach is called Cheshire II -- a standards-compliant system that can retrieve information in a variety of settings. There are many dimensions to the system including: its support of SGML; Boolean searching of servers; probabilistic ranked retrieval in the Cheshire search engine; and the search engine supports relevance feedback (Overview of Cheshire II).

Another consideration for digital libraries is autonomous citation indexing. A citation index catalogues citations that an article makes, and then links the articles with the cited works (Lawrence 1999, 67). Citation indexing is particularly helpful for the vast amount of scientific articles on the Internet. The citation index reveals relationships between articles; draws attention to corrections; and finds out how often an article is cited. An Autonomous Citation Indexing (ACI) can make a citation index from literature in electronic format.

Methodology

The author selected terms from the current literature on digital libraries to include in her mini-thesaurus. Journal articles pertaining to the indexing methods and infrastructure of digital libraries were selected from the following three electronic databases: Library Literature, Compendex, Inspec, and articles from the Internet. These three databases were chosen because they represent the disciplines of library and information science, computer science, and engineering. The results from each database were examined, and the domain-specific vocabulary was included in the thesaurus.

The thesaurus was both alphabetical and hierarchical. It maintained a controlled vocabulary of terms related to indexing methods and information infrastructure of the digital library (**See Appendix A**). Terms were separated into the following categories: **Digital Environment**; **Domain Factors**; **Indexing Methods**; **Information Infrastructure**; **Information Tools**; **Searching Process** and **Technology** (**See Appendix B**). These seven categories were selected because most of the terms the author reviewed from the literature could be easily grouped within these categories. Furthermore, the thesaurus is limited to terms that deal with the indexing methods and infrastructure of digital libraries. Hence, two of the categories were obvious selections, and the other five categories were essential to showing the relationships among all of the terms. Terms were also alphabetized letter by letter (**See Appendix C**). Maintenance of this particular thesaurus dictates its updating at the end of each month due to the progressive nature of the topic.

Discussion

After reviewing the digital library literature, and constructing a mini-thesaurus, it was obvious that there were some significant projects undertaken in an attempt to explore new concepts in the indexing of digital libraries. The creation of the mini-thesaurus highlights the need for a more unified effort among DL researchers to improve end-user access to DLs. The thesaurus contains a variety of words one can use in describing indexing methods for the digital library domain, and it reinforces the need for digital libraries to use domain-specific thesauri.

Conclusion

The goal of this research paper was to look at how professionals from the information science and engineering fields were researching the problems associated with indexing digital libraries. Marcia Bates argued that the field of information science is understaffed with researchers, and therefore cannot research many different areas at once. Even though my review of the

literature was limited to three databases, it is my belief that progress is being made on this particular concern. However, it is clear that more research must be conducted, and that the construction of subject-specific thesauri are essential if end-users are to access the wealth of information in digital libraries.

Succinctly and in essence, there is no single, unified effort to lay the foundation for research and development of digital libraries. The main purpose of this paper was to show the complexity of digital library research and development, and the need for a closer examination of indexing methods, sublanguage identification, and tools in order to facilitate quicker and more efficient access to information represented in digital library sites. The thesaurus should be considered as a prototype developed to support the above goals.

Works Cited

Bates, Marcia J. 1998. Indexing and access for digital libraries and the internet: Human, database, and domain factors. *JASIS* 49(13)(November): 1185-1205.

Borgman, Christine. 1999. What are digital libraries? Competing visions. *Information Processing and Management* 35: 227-243.

Buckland, Michael. 1999a. Mapping entry vocabulary to unfamiliar metadata vocabularies. *D-Lib Magazine* 5(1) (January). Available [Online]: http://www.dlib.org/dlib/january99/buckland/01buckland.html [May 2000].

Buckland, Michael. 1999b. Vocabulary as a central concept in library and information science. Digital Libraries: Interdisciplinary concepts, Challenges, and Opportunities: *Proceedings of the Third International Conference on Conceptions of Library and Information Science.* 3-12 (May). Available [Online]: http://www.sims.berkeley.edu/~buckland/colisvoc.htm [May 2000].

Griffin, Stephen M. 1998. Taking the initiative for digital libraries. *The Electronic Library* 16(1)(February): 24-27.

Haas, Stephanie and He, Shaoyi. 1993. Toward the automatic identification of sublanguage vocabulary. *Information Processing & Management* 29(6): 721-732.

Kessler, Jack. 1996. *Internet digital libraries: The international dimension.* Boston: Artech House, Inc.

Kramer, Ralf, Nikolai, Ralf, and Habeck, Corinna. 1997. Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on internet technologies. *International Journal on Digital Libraries* 1: 122-131.

Lawrence, Steve, Giles, C. Lee, and Bollacker, Kurt. 1999. Digital libraries and Autonomous citation indexing. *IEEE* 32(6)(June): 67-71.

Levy, David M. 2000. Digital libraries and the problem of purpose. *D-Lib Magazine.* 6(1)(January): Available [Online] http://www.dlib.org/dlib/january00/01levy.html [27 April 2000].

Lynch, Clifford and Garcia-Molina, Hector. 1996. Interoperability, scaling, and the digital libraries research agenda. *Microcomputers for Information Management: Global Internetworking for libraries* 13(2): 85-132.

Overview of Cheshire II. Available [Online] http://dli.grainger.uiuc.edu/national/berkeley/cheshire198/index.htm. [4 December 1999].

Resource indexing and discovery in a globally distributed digital library. Available [Online]: http://www2.cs.coonrell.edu/lagoze/NSF_EU/public.htm [4 December 1999].

Tennant, Roy. 1998. Digital libraries: 21st century cataloging. *Library Journal* 123(7)(April): 30-32.

Tennant, Roy. 1998. So much to digitize, so little time and money. *Library Journal* 123(13)(August): 36(2).

The NSF/DARPA/NASA Sponsored University of Michigan Digital Library Project. 1997 Available [Online]: www.si.umich.edu/UMDL/intro.html [6 December 1999].

Definition of Terms

Appendix A

Access: The ability to obtain or retrieve data or information from an information system by using indexes, thesauri, natural language searching, or other searching methods.

Automatically Extracted Keyphrases: A software technique that allows users to interact with levels of topics and collections rather than words and documents within a Keyphrase Index.

Autonomous Citation Indexing: A software program that creates a citation index from literature in electronic format through the use of World Wide Web search engines.

Boolean Searching: Searching method to improve search results by using the operators AND, OR, and NOT.

Citation Indexing: A bibliographic technique that catalogues citations an article makes, and links the articles with the cited works.

Citation Graph: A visual representation that contains all citations from a digital library through the citations' reference links.

Citation Retrieval: A retrieval technique used in connection with citation indexing and citation graphs that works by downloading papers from the World Wide Web and converting to text.

Content-Based Integration: A concept used to describe how terms may be integrated into a thesaurus by using terms retrieved from one thesaurus as an entry point to another thesaurus (Kramer 1997, 126).

Common Object Request Broker Architecture (CORBA): A standard for open distributed systems that defines ways for objects and clients to interact within a distributed environment (Kramer 1997, 126).

Descriptive Metadata: Specific metadata that identifies the author of a resource, the title, and subject heading in order to facilitate the searching and location of an item.

Digital Environment: An atmosphere concerned with the digital or electronic issues affecting a digital library.

Digitally Formatted: Any text or data converted to an electronic format so that digital libraries and information systems can make that information available to a user population geographically dispersed.

Digital Initiative: A detailed explanation of the research and development methods examined by digital library researchers.

Digital Library: A digital library is a set of electronic resources and associated technical capabilities for creating, searching, and using information by a community of users (Borgman 1999, 234).

Digital Resources: Those resources such as journal articles or audio/visual materials that have been converted to electronic format in order to be included in digital libraries.

Domain Factors: A set of issues or problems that may influence a specific domain. Those domain factors may include the following: the users' needs, the terminology important to the domain, knowledge organizations, and the structure of the language within the domain.

Domain-Specific Indexing: An examination of a particular field's vocabulary and sublanguage vocabulary to improve access to information.

English Language Indexes: An index used in Michael Buckland's work to create a bridge between unfamiliar classification or language schemes in the library and information science field and the English language

Entry Vocabulary Modules: A prototype module that aids an information seeker in the transition from familiar vocabulary to unfamiliar vocabulary in information systems.

Index Structure: The way in which words or content is structured within an index to improve search queries.

Index Structure Partitioning: The separation or division of an index structure to improve the time it takes to process a query.

Information Capture: The process of capturing information for indexing and storage.

Information Systems: A term used to describe any system that is designed for information storage and retrieval across networks.

Information Technology: The development and design of technology which improves information storage and retrieval for the end-user. An example of information technology would be the autonomous citation indexing method.

Infrastructure: The underlying framework or foundation of information tools, technology, indexing methods, and architecture which frame digital library design

Intellectual Domain: The subject-specific information or data contained within a particular sphere that makes it unique from other domains.

Keyphind: A search engine designed to support browsing in a digital library by examining broader subject topics rather than words.

Keyphrase Indexes: An index specific to the Keyphind search engine which provides a mechanism for clustering documents, refining queries, and previewing results.

Metadata: A cataloguing technique for digital libraries where structured information describes other information within a document or text.

Metadata Vocabulary: A vocabulary scheme that uses categorization codes, class numbers, indexes, and thesauri terms to describe the content within an information system.

Multidimensional Index Structure: An index structure that has many different levels or depths to the organization of the words and content within the index structure.

Multimedia Data: Digital data contained within any medium including: text documents, images, and sound (Borgman 1999, 234).

Natural Language Searching: A searching technique that permits a searcher to use any word in a search query without regard to controlled vocabulary

Ontology: A particular theory of the nature or substance of definitions. The concept is used to represent information about different thesauri databases by mapping the local scheme to a universal scheme (Kramer 1997, 127).

Searching Process: A combination of steps undertaken in order to locate and access the most relevant information possible for the end-user. The steps may include defining and limiting the subject-area one wants to examine, creating many different search queries, and using the controlled vocabulary of a particular information system.

Semantic Indexing: An indexing method that takes into account how words are used within a particular sentence or paragraph within a subject domain.

Structural Metadata: Metadata that describes how an item is structured so that a logical navigation scheme may be employed within that digital library.

Subdomain Vocabulary: Vocabulary found within a particular domain that is even more specific and precise than the general domain vocabulary.

Subject Searching: A searching technique that allows the user to formulate a statement or query about a specific topic area.

Sublanguage Analysis: The process of examining word-by-word and phrase-by-phrase sublanguage vocabulary.

Sublanguage Vocabulary: Sublanguage(SL) vocabulary is the language used in a specialized domain or topic area by the people who work in that area (Haas and He 1992, 721).

Syntactic Analysis: A method used by information scientists to examine relationships among words in support of identifying sublanguage vocabulary.

Thesaurus: An access tool that is a set of terms from a natural language vocabulary which represents the vocabulary of a particular field (Kramer 1997, 123).

Thesaurus Federation: A collection of specialized and general thesauri that allow individual thesauri to maintain their autonomy.

Vocabulary Scalability: The concept of "scaling up" systems from smaller to larger.

World Wide Web: An information technology design which is a group of interconnected INTERNET sources (Webster's New World College Dictionary 1997)

Appendix B

Digital Library Thesaurus

Hierarchical View

DIGITAL ENVIRONMENT

.digital library

..multimedia data

.digital collections

..digital resources

.digital initiative

.digitally formatted

DOMAIN FACTORS

.intellectual domain

.ontology

.semantics

.sublanguage analysis

.subdomain vocabulary

INDEXING METHODS

.autonomous citation indexing

.automatically extracted keyphrases

.citation indexing

.citation graph

.citation retrieval

.domain-specific indexing

.entry vocabulary modules

.english language indexes

.index structure

..multidimensional index structure

.keyphrase indexes

.multidimensional data indexing

.natural language indexing

.semantic indexing

.syntactic analysis

.thesaurus federation

INFORMATION INFRASTRUCTURE

.content-based integration

.information capture

.vocabulary scalability

INFORMATION TOOLS

.metadata

..descriptive metadata

..metadata vocabulary

..structural metadata

SEARCHING PROCESS

.access

.boolean searching

.natural language searching

.subject searching

TECHNOLOGY

.automated systems

.common object request broker architecture CORBA

.information technology

..world wide web

..keyphind

Appendix C

Alphabetical View

ACCESS

BT Searching Process

ACI

USE AUTONOMOUS CITATION INDEXING

AUTOMATED SYSTEMS

BT Technology

RT Information technology

AUTOMATICALLY EXTRACTED KEYPHRASES

BT Indexing methods

RT Keyphind

AUTONOMOUS CITATION INDEXING

UF ACI

BT Indexing methods

RT Citation indexing

BOOLEAN SEARCHING

BT Searching process

RT Information tools

Subject searching

CITATION GRAPH

BT Indexing methods

RT Citation indexing

Citation retrieval

CITATION INDEXING

BT Indexing methods

RT Autonomous citation indexing

Citation graph

Citation retrieval

CITATION RETRIEVAL

BT Indexing methods

RT Citation graph

Citation indexing

COMMON OBJECT REQUEST BROKER ARCHITECTURE

UF CORBA

BT Technology

RT Content based integration

CONTENT-BASED INTEGRATION

BT Information infrastructure

RT Subdomain vocabulary

Common object request broker architecture

CORBA

Use Common Object Request Broker Architecture

DESCRIPTIVE METADATA

BT Metadata

RT Information tools

Structural metadata

DIGITAL COLLECTIONS

BT Digital environment

NT Digital resources

RT Digital initiative

Digital library

DIGITAL ENVRIONMENT

NT Digital library

Digital collections

Digital initiative

Digitally formatted

RT Digital resources

Information technology

Multimedia data

Domain factors

DIGITALLY FORMATTED

BT Digital environment

DIGITAL INITIATIVE

BT Digital environment

RT Digital collections

DIGITAL LIBRARY

BT Digital environment

NT Multimedia data

RT Digital collections

Technology

Vocabulary scalability

DIGITAL RESOURCES

BT Digital collections

RT Digital environment

DOMAIN FACTORS

NT Intellectual Domain

Ontology

Semantics

Sublanguage analysis

Subdomain vocabulary

RT Domain-specific indexing

Digital environment

Indexing methods

Vocabulary scalability

DOMAIN-SPECIFIC INDEXING

BT Indexing methods

RT Domain factors

Subdomain vocabulary

ENGLISH LANGUAGE INDEXES

BT Indexing methods

RT Entry vocabulary modules

ENTRY VOCABULARY MODULES

BT Indexing methods

RT English language indexes

INDEXING METHODS

NT Autonomous citation indexing

Automatically extracted keyphrases

Citation indexing

Citation graph

Citation retrieval

Domain-specific indexing

Entry vocabulary modules

English language indexes

Keyphrase indexes

Multidimensional data indexing

Multidimensional index structure

Natural language indexing

Semantic indexing

Syntactic analysis

Thesaurus federation

RT Domain factors

INDEX STRUCTURE

BT Indexing methods

NT Index structure partitioning

Multidimensional index structure

INFORMATION CAPTURE

BT Information infrastructure

RT Information technology

INFORMATION INFRASTRUCTURE

NT Content-based integration

Information capture

Thesaurus federation

Vocabulary scalability

RT Intellectual domain

Information technology

INFORMATION TECHNOLOGY

BT Technology

NT World Wide Web

Keyphind

RT Information infrastructure

Digital environment

Information capture

INFORMATION TOOLS

NT Metadata

RT Boolean searching

Descriptive metadata

Metadata vocabulary

Natural language searching

Subject searching

Structural metadata

INTELLECTUAL DOMAIN

BT Domain factors

RT Automatically extracted keyphrases

Digital environment

Information infrastructure

KEYPHIND

BT Information technology

RT Keyphrase indexes

Technology

KEYPHRASE INDEXES

BT Indexing methods

RT Keyphind

METADATA

BT Information tools

NT Descriptive metadata

Metadata vocabulary

Structural metadata

METADATA VOCABULARY

BT Metadata

RT Information tools

MULTIDIMENSIONAL DATA INDEXING

BT Indexing methods

MULTIDIMENSIONAL INDEX STRUCTURE

BT Index structure

RT Indexing methods

MULTIMEDIA DATA

BT Digital library

RT Digital environment

NATURAL LANGUAGE INDEXING

BT Indexing methods

RT Natural language searching

NATURAL LANGUAGE SEARCHING

BT Searching process

RT Information tools

RT Natural language indexing

ONTOLOGY

BT DOMAIN FACTORS

SEARCHING PROCESS

NT Access

Boolean searching

Natural language searching

Subject searching

SEMANTICS

BT Domain factors

RT Semantic indexing

Syntactic Analysis

SEMANTIC INDEXING

BT Indexing methods

RT Semantics

Syntactic analysis

STRUCTURAL METADATA

BT Metadata

RT Descriptive metadata

Information tools

SUBDOMAIN VOCABULARY

BT Domain factors

RT Sublanguage analysis

SUBJECT SEARCHING

BT Searching process

RT Boolean searching

Information tools

SUBLANGUAGE ANALYSIS

BT Domain factors

RT Subdomain vocabulary

SYNTACTIC ANALYSIS

BT Domain factors

RT Semantics

Semantic indexing

TECHNOLOGY

NT Automated systems

Common object request broker architecture

Information technology

RT Keyphind

World Wide Web

THESAURUS FEDERATION

BT Indexing methods

VOCABULARY SCALABILITY

BT Information infrastructure

RT Digital library

Domain factors

WORLD WIDE WEB

UF WWW

BT Information technology

RT Information infrastructure

Technology

WWW

USE WORLD WIDE WEB